XNLP: A Living Survey for XAI Research in Natural Language Processing

Kun Qian
IBM Research – Almaden
San Jose, California, USA
qian.kun@ibm.com

Marina Danilevsky
IBM Research – Almaden
San Jose, California, USA
mdanile@us.ibm.com

Yannis Katsis
IBM Research – Almaden
San Jose, California, USA
yannis.katsis@ibm.com

Ban Kawas*
Facebook Research AI
USA

Erick Oduor
IBM Research – Africa
Nairobi, Kenya
ericko@ke.ibm.com

Lucian Popa
IBM Research – Almaden
San Jose, California, USA
lpopa@us.ibm.com

Yunyao Li
IBM Research – Almaden
San Jose, California, USA
yunyaoli@us.ibm.com

ABSTRACT

We present XNLP: an interactive browser-based system embodying a living survey of recent state-of-the-art research in the field of Explainable AI (XAI) within the domain of Natural Language Processing (NLP). The system visually organizes and illustrates XAI-NLP publications and distills their content to allow users to gain insights, generate ideas, and explore the field. We hope that XNLP can become a leading demonstrative example of a living survey, balancing the depth and quality of a traditional well-constructed survey paper with the collaborative dynamism of a widely available interactive tool. XNLP can be accessed at: https://xainlp2020.github.io/xainlp.

CCS CONCEPTS

• **Human-centered computing → Web-based interaction**; • **Computing methodologies → Natural language processing**.

KEYWORDS

Explainable AI, natural language processing, interactive survey

ACM Reference Format:
Kun Qian, Marina Danilevsky, Yannis Katsis, Ban Kawas, Erick Oduor, Lucian Popa, and Yunyao Li. 2021. XNLP: A Living Survey for XAI Research in Natural Language Processing. In *26th International Conference on Intelligent User Interfaces (IUI '21 Companion), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3397482.3450728

1 INTRODUCTION

One of the main goals of surveys and literature reviews[1] [3] is to capture and organize the state-of-the-art of a given field of study. A strong survey enables readers to sufficiently understand the reviewed topic, as well as effectively explore and discover new directions. After we completed a survey of the emerging field of Explainable AI (XAI) for Natural Language Processing (NLP) [4], we encountered the challenge of capturing the state of the art of a young, rapidly changing field, powered by today's unprecedented rate of publishing and speed of innovation, which has nevertheless already yielded many valuable research efforts. In the traditional format, the content of such a survey would quickly become stale, and we wished to retain the ability to dynamically and continuously update the content without compromising on the depth and quality of a well-constructed survey. We have endeavored to strike this balance in an embodiment of a living survey: an interactive browser-based system, **XNLP** . Although similar ideas has been adopted in other AI domains (e.g., [9]), to the best of our knowledge, the paradigm of a *living survey* for NLP is novel. In contrast to searchable repositories (e.g., Google Scholar[2], ACL anthology[3], Academia.edu[4]), tools (e.g., Zotero[5]) or summarization systems (e.g., [5]), a living survey should be thought of as a carefully curated online data hub that distills and synthesizes the state of the field at different levels of abstractions and from different perspectives.

XNLP is designed to reflect key characteristics of a good survey, powered by data-driven visualizations, which have been empirically shown [2] to communicate and enable more efficient ingestion of information than text alone. This also allows for better exploratory and idea generation interfaces; the identification of the current trends and gaps; and the ability for the community to add up-to-date content. This paradigm encourages the creation of

[*] Work done while working at IBM Research

[1] https://dl.acm.org/journal/csur/editorial-charter
[2] https://www.scholar.google.com
[3] https://www.aclweb.org/anthology/
[4] http://www.academia.edu/Documents/in/XAI_-_Explainable_Artificial_Intelligence
[5] https://www.zotero.org

IUI '21 Companion, April 14–17, 2021, College Station, TX, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8018-8/21/04.
https://doi.org/10.1145/3397482.3450728

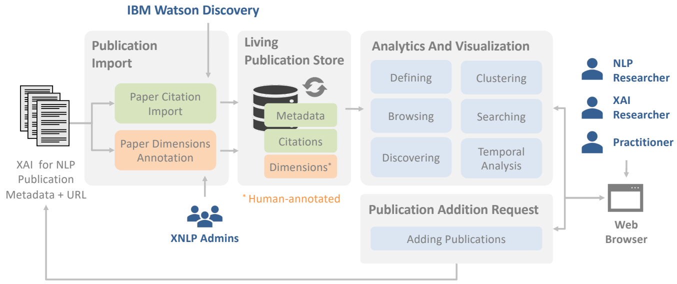


Figure 1: XNLP ’s Architecture

a community (such as other open-source communities) that would cultivate and contribute to the ongoing usefulness of **XNLP** (led by the commitment of the system creators to do the same), which, we believe, will ultimately be the key to its success.

2 DESIGN OF XNLP

XNLP is designed for different personas who may have different goals and expectations when investigating the domain. Concretely, we consider the following types of users: (1) *NLP Researcher who is new to XAI.* An NLP researcher who is not well versed in XAI is looking for a consumable perspective on the field of explainable AI within the familiar context of NLP tasks; (2) *Researcher who is new to NLP.* An XAI researcher who is not very familiar with the NLP domain is also looking for the right perspective to help her learn how to apply her general explainability expertise to a new domain; (3) *Practitioner.* This persona is mostly interested in applying the SToA XAI for NLP approaches to their own problems.
Bearing this in mind, **XNLP** is carefully designed to capture different aspects of XAI in NLP including:

- **Explanation Type**: The explanation is provided for a single prediction (local) or the entire model (global), as well as whether it is generated as part of the predictive process (self-explaining) or with additional post-processing after the prediction has been made (post-hoc).
- **Explainability Technique**: Mechanism used to generate the explanation. Common explainability techniques include feature importance, surrogate models, example-driven and provenance-based approaches, and induction.
- **Visualization Technique**: Technique used to visualize the generated explanation. Common techniques include saliency, raw declarative representations (e.g., rules), natural language, and examples.
- **Evaluation Technique**: Method used to evaluate the explanations. Common techniques include informal examination, comparison to ground truth, and human evaluation.

2.1 Publication Analytics & Visualization

As shown in Figure 1, when a PDF file of publication is imported into the system, this PDF file will be passed through two components: First, the paper’s citations are extracted using IBM Watson Discovery [6]. Second, the paper is forwarded to the **XNLP** administrators for *paper dimensions annotation.* The resulting metadata, citations, and dimensions are entered the publication store, to be used by the analytics and visualization component, described in the next section. Due to limited space, details about the criteria we adopted to collect XAI papers can be found in [4]. **XNLP** provides different views for the user to explore the XAI for NLP. Some readers may think that **XNLP** is an extension of the interactive website **XAIT** presented in [7], which to some extent is correct. However, unlike **XAIT** that provides only a single visualization for publication exploration, **XNLP** is a completely new system that enables the users to have a panorama view of the domain from different angles through different interactive visualizations. Strictly speaking, **XAIT** is only a static version of the *List view* provided by **XNLP** . We next describe the different views in **XNLP** .

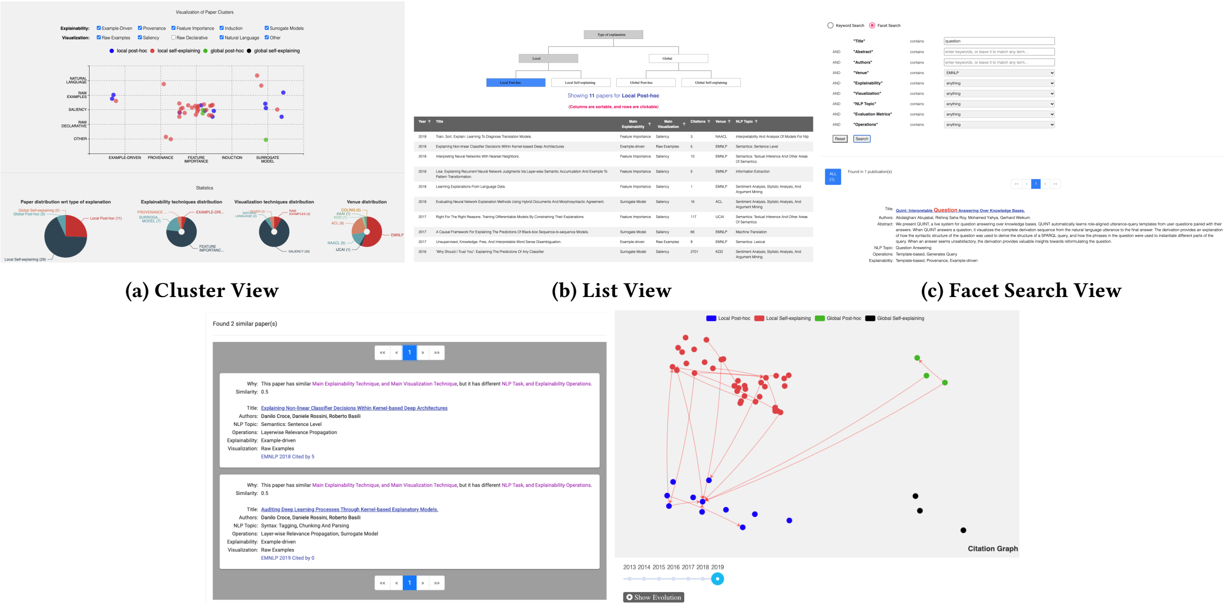
Clustering view. This view employs an interactive scatter plot to cluster papers according to three dimensions of the conceptual framework: the employed explainability techniques (shown on the x-axis), the visualization techniques (shown on the y-axis), and the explanation type (visualized through color-coding) [7]. This enables both researchers and practitioners to quickly discover the current focus of the community, which at the time of writing was on self-explaining models that use feature importance to generate explanations and visualize them through saliency techniques, such as heatmaps or highlighting (see Figure 2a). Users can also employ the filtering components on top of the scatterplot to restrict the visualization to certain types of techniques. Moreover, each point in the scatterplot is clickable and leads to a panel with additional information for the selected publication.

List view. This view displays a tabular *view* of the publications, which shows for each paper both its metadata and framework dimensions as sortable columns (Figure 2b). This can be particularly useful for researchers (both in NLP and XAI) to sort and organize the information according to their liking. Finally, this view enables users to quickly filter papers by explanation type by utilizing the corresponding clickable taxonomy at the top of the page.

Search view. While the list view provides an exploratory view of the content of the living survey, advanced users with specific goals in mind may benefit from a more guided interface. The *search view* allows users to search the content of the survey, either through conventional *keyword search* or through a more structured *faceted search.* Keyword search matches the provided keywords in all metadata and dimension fields, including title, abstract, authors, venue, NLP topics, explainability, visualization, evaluation techniques, and operations. Faceted search (Figure 2c) provides more structure by allowing users to specify search criteria for each field separately. In both cases, matched query terms are highlighted in the search results (utilizing different colors for different fields). These views can be especially useful for advanced users with a specific question in mind, such as an NLP researcher looking for work on a particular NLP topic, or an XAI researcher searching for works that employ a particular explainability or visualization technique. More importantly, each paper returned by **XNLP** also includes a *similar paper* feature. Clicking the “Find Similar Papers“ button included next to a paper displays a dialog listing similar papers. Each paper in the dialog is accompanied by both a score and a natural language

[6] https://www.ibm.com/cloud/watson-discovery

[7] Random perturbations were added to each data point to prevent publications sharing the same explainability and visualization techniques from collapsing into a single point.


Visualization of Paper Clusters
Explainability: ☑ Example-Driven ☑ Provenance ☑ Feature Importance ☑ Induction ☑ Surrogate Models
Visualization: ☑ Raw Examples ☑ Saliency ☐ Raw Declarative ☑ Natural Language ☑ Other
● local post-hoc ● local self-explaining ● global post-hoc ● global self-explaining
NATURAL LANGUAGE
RAW EXAMPLES
SALIENCY
RAW DECLARATIVE
OTHER
EXAMPLE-DRIVEN
PROVENANCE
FEATURE IMPORTANCE
INDUCTION
SURROGATE MODEL
Statistics
Paper distribution wrt type of explanation
Global Self-explaining (0)
Global Post-hoc (3)
Local Post-hoc (11)
Local Self-explaining (29)
Explainability techniques distribution
PROVENANCE ...
EXAMPLE-DRI...
SURROGA... MODEL (7)
FEATURE IMPORTANC...
Visualization techniques distribution
OTHER (2)
NATURAL LANGUAGE (2)
RAW EXAMPLES (3)
SALIENCY (32)
Venue distribution
COLING (0)
AAAI (1)
KDD (1)
ACL (8)
EMNLP
NAACL (5)
IJCAI (1)

(a) Cluster View

Showing **11** papers for **Local Post-hoc**

(Columns are sortable, and rows are clickable)

Year ↑	Title	Main Explainability ↑	Main Visualization ↑	Citations ↑	Venue ↑	NLP Topic ↑
2019	Train, Sort, Explain: Learning To Diagnose Translation Models.	Feature Importance	Saliency	3	NAACL	Interpretability And Analysis Of Models For Nlp
2018	Explaining Non-linear Classifier Decisions Within Kernel-based Deep Architectures	Example-driven	Raw Examples	5	EMNLP	Semantics: Sentence Level
2018	Interpreting Neural Networks With Nearest Neighbors.	Feature Importance	Saliency	10	EMNLP	Semantics: Textual Inference And Other Areas Of Semantics
2018	Lisa: Explaining Recurrent Neural Network Judgments Via Layer-wise Semantic Accumulation And Example To Pattern Transformation.	Feature Importance	Saliency	5	EMNLP	Information Extraction
2018	Learning Explanations From Language Data.	Feature Importance	Saliency	1	EMNLP	Sentiment Analysis, Stylistic Analysis, And Argument Mining
2018	Evaluating Neural Network Explanation Methods Using Hybrid Documents And Morphosyntactic Agreement.	Surrogate Model	Saliency	16	ACL	Sentiment Analysis, Stylistic Analysis, And Argument Mining
2017	Right For The Right Reasons: Training Differentiable Models By Constraining Their Explanations	Feature Importance	Saliency	117	IJCAI	Semantics: Textual Inference And Other Areas Of Semantics
2017	A Causal Framework For Explaining The Predictions Of Black-box Sequence-to-sequence Models.	Surrogate Model	Saliency	66	EMNLP	Machine Translation
2017	Unsupervised, Knowledge- Free, And Interpretable Word Sense Disambiguation.	Example-driven	Raw Examples	9	EMNLP	Semantics: Lexical
2016	"Why Should I Trust You": Explaining The Predictions Of Any Classifier.	Surrogate Model	Saliency	2701	KDD	Sentiment Analysis, Stylistic Analysis, And Argument Mining

(b) List View

○ Keyword Search ● Facet Search

	"Title"	contains	question
AND	"Abstract"	contains	enter keywords, or leave it to match any term...
AND	"Authors"	contains	enter keywords, or leave it to match any term...
AND	"Venue"	contains	EMNLP
AND	"Explainability"	contains	anything
AND	"Visualization"	contains	anything
AND	"NLP Topic"	contains	anything
AND	"Evaluation Metrics"	contains	anything
AND	"Operations"	contains	anything

Found in 1 publication(s)

Title: Quint: Interpretable Question Answering Over Knowledge Bases.
Authors: Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, Gerhard Weikum
Abstract: We present QUINT, a live system for question answering over knowledge bases. QUINT automatically learns role-aligned utterance-query templates from user questions paired with their answers. When QUINT answers a question, it visualizes the complete derivation sequence from the natural language utterance to the final answer. The derivation provides an explanation of how the syntactic structure of the question was used to derive the structure of a SPARQL query, and how the phrases in the question were used to instantiate different parts of the query. When an answer seems unsatisfactory, the derivation provides valuable insights towards reformulating the question.
NLP Topic: Question Answering
Operations: Template-based, Generates Query
Explainability: Template-based, Provenance, Example-driven

(c) Facet Search View

Found 2 similar paper(s)

Why: This paper has similar Main Explainability Technique, and Main Visualization Technique, but it has different NLP Task, and Explainability Operations.
Similarity: 0.5

Title: Explaining Non-linear Classifier Decisions Within Kernel-based Deep Architectures
Authors: Danilo Croce, Daniele Rossini, Roberto Basili
NLP Topic: Semantics: Sentence Level
Operations: Layerwise Relevance Propagation
Explainability: Example-driven
Visualization: Raw Examples
EMNLP 2018 Cited by 5

Why: This paper has similar Main Explainability Technique, and Main Visualization Technique, but it has different NLP Task, and Explainability Operations.
Similarity: 0.5

Title: Auditing Deep Learning Processes Through Kernel-based Explanatory Models.
Authors: Danilo Croce, Daniele Rossini, Roberto Basili
NLP Topic: Syntax: Tagging, Chunking And Parsing
Operations: Layer-wise Relevance Propagation, Surrogate Model
Explainability: Example-driven
Visualization: Raw Examples
EMNLP 2019 Cited by 0

(d) Similar Papers Dialog

🟦 Local Post-hoc 🟥 Local Self-explaining 🟩 Global Post-hoc ⬛ Global Self-explaining
Citation Graph
2013 2014 2015 2016 2017 2018 2019
Show Evolution

(e) Citation Graph View

Figure 2: Screenshots of different publication views offered by XNLP

explanation of its similarity (Figure 2d). The similarity between two papers is calculated based on their explainability techniques, visualization techniques, NLP topics, and operations and the natural language explanation is produced through a template-based approach. Although a relatively simple approach, it is nonetheless more effective than conventional similar publication discovery methods, which base the similarity score solely on standard metadata (such as co-authors). We are also exploring more advanced similarity techniques (e.g., embedding-based similarity) in the future.

Citation graph view. While the aforementioned functionality allows users to easily browse and find relevant publications, it does not make clear which publications the community considers to be important, or how the field has evolved over time. To answer these questions, **XNLP** analyzes paper citations and presents them in a citation graph, drawing inspiration from related work in analyzing scientific activity [1, 6]. The resulting *citation graph view* (see Figure 2e) allows NLP and XAI researchers to easily spot influential publications, such as [8], one of the well-known pioneers in the field. Finally, to communicate the evolution of the field, **XNLP** extends the citation graph by showing its evolution over time, which can be important for understanding the achievements and interests of the community. For instance, at the time of this writing, a clear trend of increasing publication numbers in recent years can be observed.

3 DEMO EXPERIENCE

Consider one of our personas, an NLP researcher who is not very familiar with XAI. She works on Question Answering (QA) tasks. Reading existing related surveys and tutorials gives her background knowledge on explainability, but she is still not clear on how to apply it to her own research. Upon landing on the home page of **XNLP** , she can peruse the Definitions page to gain some background knowledge on the conceptual framework and terminology, as well as get pointers to example papers. Ready to focus on learning more concretely about her own topic, our researcher clicks over to the Faceted Search (Figure 2c) and selects *Question Answering* for the NLP Task. Scrolling through the focused results, she finds significant variations across all four dimensions over a relatively small number of papers, indicating that there are many different ways to approach XAI for QA, and possibly suggesting some new research directions to her.

REFERENCES

[1] Yuan An, Jeannette Janssen, and Evangelos E Milios. 2004. Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems* 6, 6 (2004), 664–678.

[2] Eliza Bobek and Barbara Tversky. 2016. Creating visual explanations improves learning. *Cognitive Research: Principles and Implications* 1, 1 (2016), 27.

[3] Jan vom Brocke, Alexander Simons, Björn Niehaves, Kai Riemer, Ralf Plattfaut, and Anne Cleven. 2009. Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process.

[4] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. *AACL-IJCNLP 2020.*

[5] Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. [n.d.]. A Summarization System for Scientific Documents. In *EMNLP 2019.* 211–216. https://www.aclweb.org/anthology/D19-3036

[6] Morgan R Frank, Dashun Wang, Manuel Cebrian, and Iyad Rahwan. 2019. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence* 1, 2 (2019), 79–85.

[7] Erick Oduor, Kun Qian, Yunyao Li, and Lucian Popa. 2020. XAIT: An Interactive Website for Explainable AI for Text. In *ACM IUI'20.* New York, NY, USA, 120–121. https://doi.org/10.1145/3379336.3381468

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *KDD.* 1135–1144.

[9] Qianwen Wang, Jun Yuan, Shuxin Chen, Hang Su, Huamin Qu, and Shixia Liu. 2019. Visual genealogy of deep neural networks. *IEEE transactions on visualization and computer graphics* 26, 11 (2019), 3340–3352.